Validation of a loudspeaker-based room auralization system using speech intelligibility measures

Favrot, Sylvain Emmanuel; Buchholz, Jörg

Published in:
Audio Engineering Society Convention Papers

Publication date:
2009

Document Version
Publisher's PDF, also known as Version of record

Audio Engineering Society

Convention Paper 7763

Presented at the 126th Convention
2009 May 7–10 Munich, Germany

Validation of a loudspeaker-based room auralization system using speech intelligibility measures

Sylvain Favrot[1], Jörg M. Buchholz[1]

[1] *Centre for Applied Hearing Research, Technical University of Denmark, Kgs. Lyngby, Denmark*

Correspondence should be addressed to Sylvain Favrot (sf@elektro.dtu.dk)

ABSTRACT

A novel loudspeaker-based room auralization (LoRA) system has been proposed to generate versatile and realistic virtual auditory environments (VAEs) for investigating human auditory perception. This system efficiently combines modern room acoustic models with loudspeaker auralization using either single loudspeaker or high-order Ambisonics (HOA) auralization. The LoRA signal processing of the direct sound and the early reflections was investigated by measuring the speech intelligibility enhancement by early reflections in diffuse background noise. Danish sentences were simulated in a classroom and the direct sound and each early reflection were either auralized with a single loudspeaker, HOA or first-order Ambisonics. Results indicated that (i) absolute intelligibility scores are significantly dependent on the reproduced technique and that (ii) early reflections reproduced with HOA provide a similar benefit on intelligibility as when reproduced with a single loudspeaker. It is concluded that speech intelligibility experiments can be carried out with the LoRA system either with the single loudspeaker or HOA technique.

1. INTRODUCTION

Recently, a novel loudspeaker-based room auralisation (LoRA) system has been proposed [1], which aims at generating fully controllable and highly realistic virtual auditory environments (VAEs) that are suitable for investigating human auditory perception as well as assessing and optimizing the performance of modern hearing devices.

The LoRA system effectively combines state-of-the art acoustic room models and loudspeaker-based auralization. Each component of the discrete part of the room's response (i.e., the direct sound and the

early reflections) is auralized individually as a single source. Different reproduction techniques can be chosen to auralize these discrete components using either single loudspeakers (the closest from the component's incoming direction) or Ambisonics (first-order [2] or high-order [3]). The reproduction of the (late) diffuse reverberation part is realized by multiplying (directional) intensity envelopes of the room's response with noise that is uncorrelated across loudspeakers.

When a large number of loudspeakers is available, then the single loudspeaker technique is the most accurate one for reproducing the direct sound and the early reflections. Ambisonics allows the reproduction of sound events from any direction, whereby the localization accuracy depends on the applied Ambisonic order which itself depends on the available number of loudspeakers [3]. Ambisonics reproduces the sound field accurately inside a sweet spot (e.g., in the center of a loudspeaker array) up to a certain frequency f_{max}. This frequency f_{max} and the size of the sweet spot are determined by the applied Ambisonic order (e.g., $f_{max} = 2.2$ kHz for forth-order Ambisonics within a sweet spot of 20 cm diameter). Apart from reproducing simulated reflections, this technique can be used to reproduce a whole auditory scene which has been recorded either with a standard soundfield microphone [2] (first-order Ambisonics) or with a more advanced microphone array (higher-order Ambisonics, HOA) [4]. Ambisonics is also suitable for reproducing moving sound sources as it allows for smooth source direction changes. Hence, although the single loudspeaker technique might in principle provide the best overall quality, Ambisonics represents a more versatile method and might be superior when only a limited number of loudspeakers are available.

An objective evaluation using different room acoustic measures showed that the spectral, temporal and spatial aspects of the room's response are preserved by the LoRA processing [1]. Since the LoRA system is primarily designed for auditory perception research, a subjective evaluation of the system is needed. This study presents a speech intelligibility measure-based subjective evaluation of the auralization techniques used in the LoRA system. The specific aim was to assess the influence of single loudspeaker, first-order Ambisonics and HOA auralization of the direct sound and early reflections on speech intelligibility scores. The results allow conclusions on the applicability of (higher-order) Ambisonics in VAEs (such as the LoRA system) for speech perception research and moreover, have direct implications for simulating moving speech sources as well as presenting speech signals recorded in real environments.

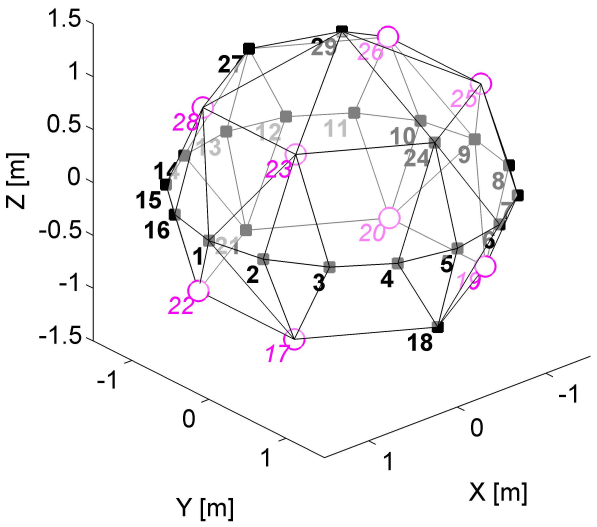

Fig. 1: Loudspeaker positions of the array used in the experiments. The large open circles indicate the loudspeakers used for first-order Ambisonics auralization. In the single loudspeaker auralization condition, the direct sound was played with loudspeaker '1' only.

A study by Shirley et al. [5] has shown that speech intelligibility measures can produce significantly different results when speech is presented by a single loudspeaker than when mixed via a stereo loudspeaker pair. Results might expected to be different here from those of Shirley et al. because (i) first-order Ambisonic and especially high-order Ambisonic are expected to provide a more accurate reproduction of the sound field of a single sound source than stereo mixing and (ii) listeners sensitivity to the direct sound image is limited here due to the presence of (simulated) room reflections.

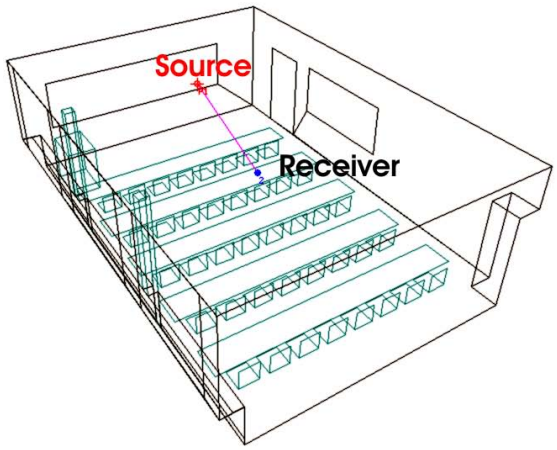

Fig. 2: Position of the source and the receiver in the 3D model of a classroom (40 seats, 170 m^3). The listener was located at a distance of 3.5 m from the talker.

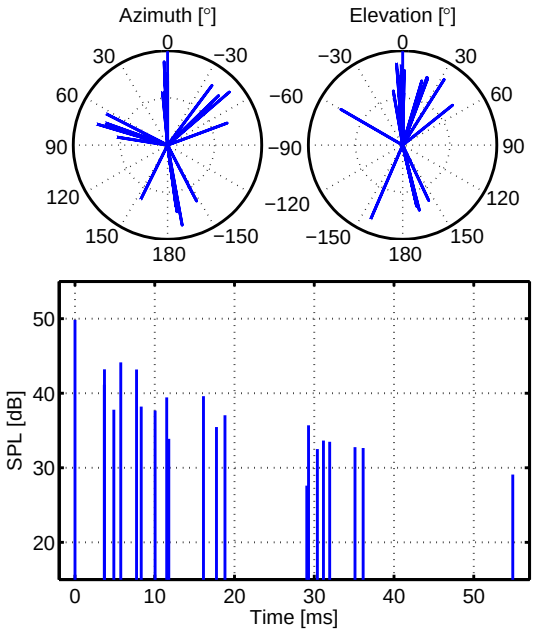

Fig. 3: Reflectogram for the source and receiver configuration in the classroom showed in Fig. 2). The directional characteristic (azimuth and elevation) is indicated in the polar plots and the temporal behavior is shown in the lower panel.

2. METHOD

The influence of auralization technique on speech intelligibility in virtual environments was investigated by applying a method inspired from Bradley et al. [6]. In order to separately evaluate the effect of the auralization technique on the direct sound and early reflections, speech intelligibility scores were measured as a function of direct sound level and early reflection level.

The experiment took place in an acoustically damped room where a 3-D array of 29 loudspeakers was used (see Fig. 1). Nine normal hearing persons participated in the experiment.

2.1. Stimuli

2.1.1. Room simulation

A classroom was simulated with ODEON (see Fig. 2), a room acoustic modeling software, where a source ('talker') and a receiver ('listener') position were defined. A reflectogram (i.e. the direction, latency and attenuation of the direct sound (DS) and early reflections (ER), see Fig. 3) was obtained for this source-receiver configuration.

The direction of each individual component of the reflectogram was manipulated to match the direction of the closest loudspeaker present in the listening room (see Fig. 1). This was done to ensure that the single loudspeaker and Ambisonic auralization represent reflections from the same direction. The 'talker' was facing the 'source' such that the direct sound came from the frontal direction (0° azimuth, 0° elevation).

The reflectogram data were then processed separately for the direct sound and the early reflections with the LoRA toolbox. Each discrete component was treated as a source being reproduced either with a single loudspeaker (labeled as '0') , with fourth-order Ambisonics (labeled as '4') or with standard first-order Ambisonics (labeled as '1'). For the Ambisonic reproduction, a 'basic' decoding scheme was used for frequencies up to 2.8 kHz for fourth-order Ambisonics and up to 707 Hz for first-order Ambisonics. Above this frequency, the 'max r_E' decoding method was used in order to focus the energy in the expected direction [3]. Only 8 loudspeakers, indicated with large open circles in Fig. 1, were used for first-order Ambisonic auralization in order

to limit coloration effects.

Since speech intelligibility in diffuse noise was measured here, the late reverberation part of the room's response was not reproduced. The presence of late reverberation is expected to deteriorate speech intelligibility, however the level of the late reverberation in the classroom was low relatively to the level of the diffuse noise.

For the source-receiver configuration in the considered room, the obtained multichannel room impulse response (mRIR) was about 55 ms long and the level of the total mRIR (DS+ER) was 4 dB higher than the level of the direct sound (DS) alone.

The level of the direct sound and/or of the early reflections in the mRIR was varied to obtained different signal-to-noise ratios (SNRs).

2.1.2. Speech corpus and speech-shaped noise

The speech corpus consisted of the Dantale II Danish Hagerman sentences [7]; each sentence containing five words following the structure: 'Name' + 'Verb' + 'Number' + 'Adjective' + 'Noun'. The sentences were auralized by convolving them with the classroom mRIR reproduced with different techniques.

Diffuse speech-shaped noise was obtained from cutting the monaural speech-shaped noise track from the Dantale II material in 29 noise signals. These uncorrelated noise signals were played simultaneously via the 29 loudspeakers. The diffuse noise started 1 s before the sentence with a cosine-shaped ramp of 0.6 s and stopped 0.5 s after the sentence with a decreasing cosine-shaped ramp of 0.3 s. Diffuse noise was played at a fixed level of 60 dB SPL.

2.1.3. Calibration

Loudspeaker equalization was performed on the loudspeaker array to flatten each of the loudspeaker frequency responses recorded at the center of the array. In order to calibrate the system, speech-shaped noise convolved with the DS-only and the ER-only for the different conditions were recorded with an omnidirectional microphone at different positions within the sweet spot (7.5 cm around the center of the loudspeaker array).

The SPL of the recordings for the DS-only auralization with the three different techniques showed discrepancies with the simulated level ranging from -1 dB to +2.5 dB. A level adjustment was then performed to ensure identical SPL independent of the direct sound reproduction technique. A similar adjustment was realized for the ER-only responses which initially showed discrepancies from 2.2 dB to 3.6 dB. When the whole mRIRs (with the adjusted DS and ER levels) were used, the obtained SPLs varied only from 0 to +1.5 dB compared to the simulated SPLs. Different dummy-head recordings in the sweet spot as well as slightly off-center confirmed the applicability of the level-adjustment method (see Fig. 4 for the sweet spot recordings).

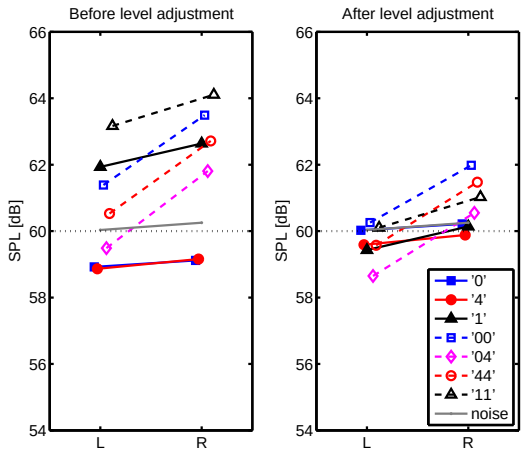

Fig. 4: Binaural levels before and after level adjustment for the seven conditions. Each condition was auralized at 60 dB SPL. The measured SPL of the diffuse speech-shaped noise is also plotted.

2.1.4. Discrimination task

A preliminary test was carried out to determine if listeners could discriminate short sentences presented with the different auralization techniques. DS-only and DS+ER mRIRs auralized with the three previously mentioned techniques were used for this test. Diffuse noise was not played simultaneously during the short sentences. Results showed that all techniques could be clearly discriminated. This was a prerequisite for the speech intelligibility experiment as non discriminable stimuli would have provided similar intelligibility scores.

2.2. Procedure

First, the speech reception threshold (SRT) was de-

rived with an adaptive method where the whole mRIR (DS+ER) level was varied relative to the fixed level of the diffuse speech-shaped noise (60 dB SPL). The SRT corresponded to the signal-to-noise ratio (SNR) at 50 % intelligibility. The DS reference level was defined as the level of the direct sound in the whole mRIR at the SRT level. The mRIR used in this part of the experiment was obtained by the single loudspeaker technique for both the DS and the ER (condition '00'). The SRT was measured for each subject with two lists of 10 sentences.

Second, intelligibility scores were measured for fixed SNRs of SRT -2, 0, +2, and +4 dB by using only the direct sound and varying its level. Measures were taken for the direct sound auralized by (i) a single loudspeaker, (ii) forth-order HOA and (iii) first-order Ambisonics (conditions '0', '4' and '1' respectively).

Third, the direct sound level was kept at the previously derived DS reference level and SNRs of SRT -2, 0, +2, and +4 dB were obtained by adding early reflections to the direct sound and varying the level of these early reflections. Four conditions were measured for this part of this experiment: the same technique was used for the DS and ER (conditions '00', '44' and '11') and another condition consisted in auralizing the DS with a single loudspeaker and the ER with forth-order Ambisonics (condition '04'). For the second and third part of this experiment, a list of 10 sentences was used to determined the intelligibility scores.

In each part of the experiment, after each sentence was played with diffuse noise, subjects were asked to select the five words s/he had heard on a touch screen (see Fig. 5). There were ten possible choices for each word plus a "I don't know" (labeled as "?") one which, when selected, randomly choose a word from the list of 10 words. The "I don't know" button was present to not force the subject to randomly pick a word if it was not heard. The selection screen was shown only after the sentence was played. After the selection, the subject pressed the "OK" button to play the next sentence.

Each test-subject participated in a training phase which consisted of 4 repetitions of the first part of the experiment (80 sentences). After the first part of the experiment, intelligibility scores for the different

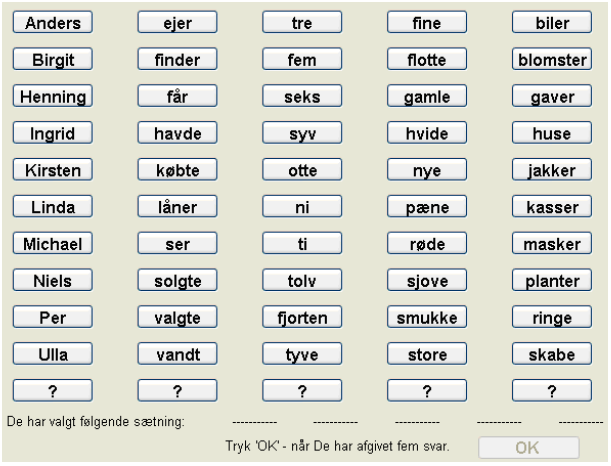

Fig. 5: Touch screen caption for the sentence selection.

SNRs and conditions were measured in a random order. The entire experiment was carried out for each subject in two sessions of one hour and a half, including breaks.

3. RESULTS

In the first part of the experiment, speech reception thresholds were measured between -13 and -10.3 dB for the nine test-subjects. These values are substantially lower than the -8.4 dB normally observed for normal hearing subjects with the classic Dantale II test. This difference might be explained by the playback method, i.e. using loudspeakers instead of headphones, and the experimental procedures, i.e. applying a user interface with a restricted choice of words rather than the subject telling the operator what s/he had heard.

Fig. 6 and Fig. 7 show the inter-subject mean intelligibility scores for the DS-only conditions and the DS+ER conditions respectively. The corresponding standard deviations are indicated by error-bars. Results from the nine test-subjects were analyzed for these two groups of conditions. To assess the statistical significance of the comparison of any two conditions, a paired t-test was performed.

3.1. Modeling the data

As expected, mean intelligibility scores plotted over SNR exhibited a psychometric function. In order to quantify the measured psychometric functions, a

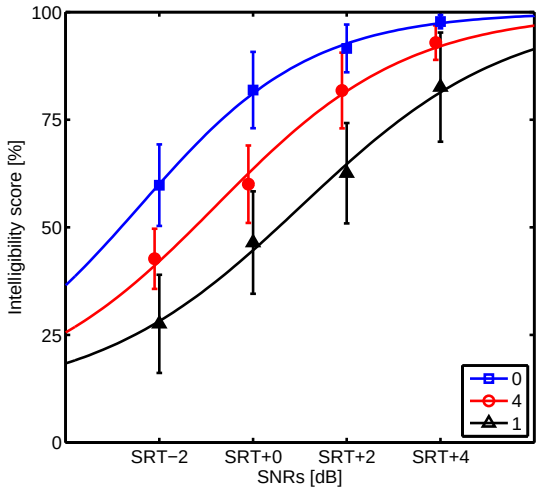

Fig. 6: Direct sound only mean speech intelligibility scores (markers) with ± 1 standard deviation. Solid lines represent the fitted sigmoid function.

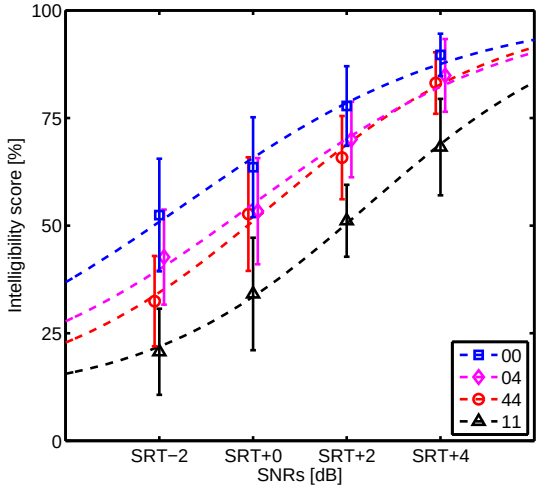

Fig. 7: Direct sound and early reflections mean speech intelligibility scores (markers) with ± 1 standard deviation. Dashed lines represent the fitted sigmoid function.

sigmoid function $P(L)$ was fitted to the data with the parameter L_{55} (threshold, SNR L at 55 % percent correct) and s_{55} (slope at the inflection point, $L = L_{55}$) for each condition according to the following formula [8]:

$$P(L) = \frac{1-\alpha}{1+exp(4 \cdot s_{55} \cdot (L_{55}+L))} + \alpha. \quad (1)$$

The chance level α was here 10 % since there were 10 possible answers.

The fitted parameters can be found in Table 1 and the corresponding sigmoid functions were plotted in Fig. 6 (solid lines) and Fig. 7 (dotted lines). The sigmoid functions fitted the measured data with an RMS error smaller than 2 %. Theses parameters characterize the significant increase of intelligibility score with increasing SNR and were analyzed for all the considered conditions.

Conditions	0	4	1	00	44	11
L_{55} [dB]	-2.4	-0.7	1.0	-1.5	0.5	2.5
s_{55} [%/dB]	13.7	12.2	11.3	8.4	10.1	10.5
error [%]	0.7	1.7	1.6	1.8	1.8	0.8

Table 1: Fitted parameters of the sigmoid function for each group.

3.2. Effect of the reproduction technique for the direct sound only conditions

For the direct sound only conditions (see Fig. Fig. 6 and Table 1), the intelligibility scores are significantly dependent on the auralization method. Highest scores are observed for the single loudspeaker technique and lowest scores for first order Ambisonics. The difference is mainly due to a shift in the L_{55} threshold, indicating a shift in effective SNR: a threshold shift of 1.7 dB is observed from single loudspeaker ('0') to forth-order HOA ('4') and a shift of 3.4 dB from single loudspeaker to first-order Ambisonics ('1').

Moreover, the psychometric functions for the '0' condition are steeper at the inflection point than for the Ambisonic ones: the s_{55} slope decreases by 1.5 %/dB for HOA and by 2.4 %/dB for the first-order Ambisonics one. These small variations indicate that, with Ambisonics (especially at first-order) slightly larger SNRs than with the single loudspeaker technique are required to increase intelligibility scores.

3.3. Effect of the auralization technique on the reproduction of the whole response

For the DS+ER conditions (Fig. 7 and Table 1), the

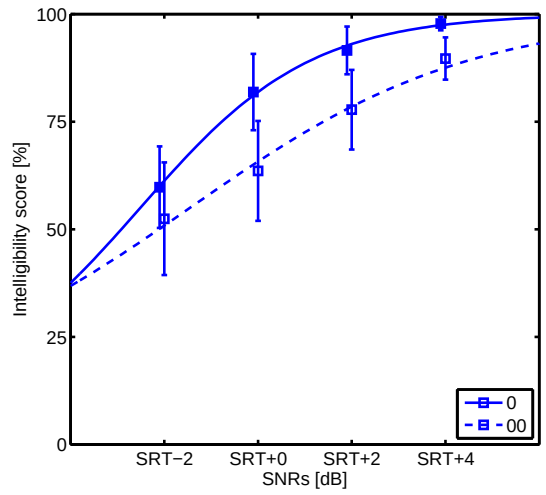

Fig. 8: Mean speech intelligibility scores (markers) and psychometric curves for DS-only and DS+ER conditions auralized with single loudspeaker.

intelligibility scores show a similar significant dependency on auralization technique as observed for the direct sound only conditions (section 3.2). For HOA auralization a L_{55} threshold shift of 2.0 dB is needed to obtain similar intelligibility scores than with the single loudspeaker technique. For first-order Ambisonic a 4 dB threshold shift is required to match the scores obtained with the single loudspeaker auralization.

The slope of the psychometric functions are slightly steeper with Ambisonics than with the single loudspeaker technique. These slope variations with the reproduction technique are the opposite of the ones observed for the DS-only conditions. However, these variations are rather small and might thus be disregarded.

Since condition '04' and '44' showed no significant score difference, the fitted parameters for condition '04' were not considered in this analysis.

3.4. Effect of the addition of early reflections

Although overall intelligibility is modified by the playback technique, the addition of reflections has a similar effect for all playback techniques.

Comparing the intelligibility scores measured in the DS-only condition (Fig. 6) and DS+ER conditions (Fig. 7), it can be observed that the increase in intelligibility is larger when adding direct sound energy than when adding reflection energy. Accordingly, the $s55$ slopes of the psychometric functions (Table 1) are shallower for the DS+ER conditions than for the DS-only conditions. The effect is highlighted in Fig. 8 for the single loudspeaker technique where the experimental data is replotted for condition '0' and condition '00'. The decrease in s_{55} slope is slightly more pronounced for the single loudspeaker technique (-5.3 %/dB) than for HOA (-2.1 %/dB) and first-order Ambisonics (-0.8 %/dB).

The L_{55} thresholds increase for DS+ER conditions compared to DS-only conditions with slightly variation for different reproduction technique. A threshold increase of 1.0 dB was observed for the single loudspeaker technique, 1.2 dB for HOA and 1.4 dB for first-order Ambisonics (see Table 1).

4. DISCUSSION

The fitted psychometric functions showed that in the present experiment the absolute intelligibility scores were dependent on the auralization technique for both the DS-only and DS+ER conditions. The dummy-head recording levels for the DS-only conditions were 0.3 dB smaller for the two Ambisonics techniques compared to the single loudspeaker auralization (see Fig. 4). Since thresholds shifts of the psychometric functions for the different auralization techniques were in the order of 1.7-3.4 dB (Fig. 6 and Table 1), they can not be solely explained by the DS level calibration. For the DS+ER conditions, the dummy-head recording levels showed a decrease of about 0.5 dB also failing to account for the shifts of the psychometric function of 2-4 dB (see Fig. 7).

The decrease in intelligibility was probably due to the imperfect sound field reconstruction with Ambisonics. This typically leads to degraded spatial cues and coloration due to comb-filtering effects. These effects are even more pronounced when the Ambisonic order is low, which is here reflected by the lower scores for first-order Ambisonics than for forth-order Ambisonics. These detrimental effects for an anechoic source (DS-only) and for a source in a reverberant environment (DS+ER) did also lower the intelligibility of forth-order Ambisonics compared to the single loudspeaker technique. This observation is in line with Shirley's study [5] where a degraded

speech intelligibility was measured when using a stereo phantom image compared to a single loudspeaker due to the cross-talk caused by the phantom image. With Ambisonics, the cross-talk can be described with the energy ratio or spatial spread r_E [3] which is increasing with Ambisonics order but does not reach the value of 1 which would be obtained for the reference (single loudspeaker).

The SNR increase by raising direct sound level versus addition of early reflections led to a decrease in the slope of the psychometric function as well as a threshold increase for all the considered reproduction techniques. This can be interpreted as the consequence of a temporal and spatial spread of energy when early reflections are added. The benefit of early reflections in speech intelligibility has been pointed out in numerous studies ([9] and [10]). However, in the study by Bradley et al. [6], the benefit of the early reflections was of the same order as for an increase of the direct sound level only. This finding was not observed here and might be due to a difference in the employed speech test as well as the considered stimuli.

5. CONCLUSION

This study investigated the impact of the auralization method on speech intelligibility. It was found that the overall intelligibility threshold (SNR at 55 % word correct) was lower when using a single loudspeaker technique than when using Ambisonics. This threshold also increased when the Ambisonic order decreased. Moreover, the addition of early reflections increased the intelligibility to a lesser extent than when the direct sound alone was raised, resulting in psychometric functions with shallower slopes. However, the addition of early reflections induced a similar threshold shift for all the considered reproduction techniques.

It can be concluded that speech intelligibility experiments can be run with the LoRA system with either the single loudspeaker or HOA technique as the reproduced early reflections provide the same benefit on intelligibility. However, intelligibility scores need to be equalized for the individual auralization method by a simple SNR shift. This encourages the use of HOA microphone arrays to record complex auditory scenes for speech intelligibility experiments. However, further evaluation is required to investigate the effect of the physical properties of such microphone arrays on the captured auditory scene.

6. ACKNOWLEDGMENT

The authors thank Iris Arweiler and Torsten Dau for valuable discussions. The study was supported by a stipend from the Technical University of Denmark.

7. REFERENCES

[1] S. Favrot and J. Buchholz, "LoRA - A loudspeaker-based room auralisation system ," Acta Acustica, submitted, 2009.

[2] M. A. Gerzon, "Periphony - with-height sound reproduction," Journal of the Audio Engineering Society, **21**, 1973.

[3] J. Daniel, "Representation de champs acoustiques, application a la transmission et a la reproduction de scenes sonores complexes dans un contexte multimedia (in french)," PhD thesis, universite Paris 6, France, 2000.

[4] S. Moreau, J. Daniel and S. Bertet, "3D Sound Field Recording with Higher Order Ambisonics," presented at the AES 120th convention, Paris, France, 2006.

[5] B. Shirley, P. Kendrick and C. Churchill, "The effect of stereo crosstalk on intelligibility: comparison of a phantom stereo image and a central loudspeaker source," Journal of the Audio Engineering Society, **55**, 2007.

[6] J. S. Bradley and H. Sato, "On the importance of early reflections for speech in rooms," Journal of the Acoustical Society of America, **113**, 2003.

[7] K. Wagener, J. L. Josvassen and R. Ardenkjaer, "Design, optimization and evaluation of a Danish sentence test in noise," International Journal of Audiology, **42**, 2003.

[8] T. Brand and B. Kollmeier, "Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests," Journal of the Acoustical Society of America, **111**, 2002.

[9] J. P. A. Lochner and J. F Burger, "The influence of reflections on auditorium acoustics," Journal of Sound and Vibration, **42**, 1964.

[10] A. K. Nabelek and L. Robinette, "Influence of the precedence effect on word identification by normally hearing and hearing-impaired subjects," Journal of the Acoustical Society of America, **63**, 1978.